UDC: 004.89 https://doi.org/10.15407/jai2025.01.141

I. Olenych[1], **D. Demchyk**[2], **S. Babiak**[3], **O. Futey**[4]
[1,2,3,4]Ivan Franko National University of Lviv, Ukraine
50 Dragomanov Street, Lviv 79005
[1]igor.olenych@lnu.edu.ua
[2]d.demchik15@gmail.com
[3]frageks@gmail.com
[4]oleksandr.futey@lnu.edu.ua
[1]https://orcid.org/0000-0002-6642-0222
[2]https://orcid.org/0009-0000-0495-2939
[3]https://orcid.org/0009-0002-8726-2742
[4]https://orcid.org/0000-0002-6491-1669

AIR POLLUTION PREDICTION USING MACHINE LEARNING

Abstract. Prediction of air pollution with particulate matter is a critically important task for developing effective strategies to improve the environmental situation. Despite the large number of predictive machine learning models, insufficient attention has been paid to investigating the effectiveness of pollution prediction in different ranges of microparticle concentrations. The paper proposes models for forecasting atmospheric pollution with particulate matter up to 2.5 microns in size (PM2.5) based on the Long Short-Term Memory (LSTM), Extreme Gradient Boosting (XGBoost), and Random Forest algorithms taking into account meteorological and spatio-temporal data obtained by the developed air quality monitoring system. Particular attention was focused on studying the dependence of forecasting accuracy on the level of atmospheric pollution.

It was found that the proposed models successfully predict the PM2.5 content in the air at low and medium levels of pollution but underestimate the predicted values as their concentration increases. Based on the analysis of the concentration dependences of absolute and relative errors, it was found that the Random Forest method demonstrates the highest prediction accuracy in a wide range of the PM2.5 concentration with a relative error of 6–9 % despite deviations for some peak values. Models based on the XGBoost and LSTM methods are characterized by errors of 9–11 and 11–14 %, respectively. A decrease in forecast accuracy and a significant increase in the variance of predicted values were found with an increase in the concentration of the particulate matter in the air. The LSTM method demonstrates the worst results for high levels of air pollution. The decrease in the effectiveness of predictive models with increasing atmospheric pollution may be due to the small number of records with a high concentration of particulate matter in the dataset and the random appearance of additional pollution sources unrelated to meteorological conditions and spatio-temporal characteristics. An integral assessment of the accuracy of the developed models using the metrics Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination R^2 confirms the high efficiency of predicting the PM2.5 concentration in the air.

Keywords: air pollution, forecasting, model efficiency, machine learning, artificial neural networks.

Introduction

The rapid growth of urbanization and industrialization is the source of many environmental problems in the modern world. Among the most important problems of urban agglomerations and industrial regions is the deterioration of air quality [1]. In particular, atmospheric pollution with particulate matter up to 2.5 microns in size (PM2.5) is not only a threat to human health but also has a fundamental impact on global climate change [2-4]. Such particles can spread quickly and easily over long distances and remain in the atmosphere for a long time due to their small size. In addition to the direct threat to the human respiratory and cardiovascular systems, PM2.5 particles can adsorb toxic substances on their surface. As a result of this combination, the negative impact of particulate matter may increase [5], which makes it impossible to accurately determine the safe concentration of pollutants and necessitates comprehensive air quality monitoring and the all-round analysis of measurement results.

An important component of the intelligent environmental monitoring system is the pre-processing of sensor data, which includes their aggregation, detection of omissions and erroneous records as well as the removal of duplications. In addition, algorithms for the detection and identification of outliers in measured data are of great importance [6,7]. The outlier processing makes it possible to develop air pollution models that take into account anomalies in the time series of sensor data. Big data and artificial

intelligence (AI) technologies provide the necessary tools to identify patterns and create such models [8].

Estimating the concentration of the PM2.5 particles and predicting the level of air pollution are critically important tasks for the development of effective risk management strategies and the formation of policies to improve the environmental situation [9,10]. Air quality forecasting can provide early warning of potential increases in pollutant concentrations to effectively prevent pollution or minimize its effects.

Machine learning methods that are gaining popularity as powerful tools for forecasting in various fields provide new opportunities for analyzing air quality [11,12]. However, analyzing multidimensional data with complex spatial and temporal dependencies influenced by many external factors poses a serious challenge for machine learning traditional approaches [13]. More accurate and reliable predictions are provided by deep learning models that can identify multiple correlations and patterns in air pollution data [14]. Such models usually take into account various auxiliary information such as meteorological data (e.g., air temperature and relative humidity, wind speed, precipitation), geographical characteristics, pollution sources, etc. [15-17].

Although many models with diverse architectures have been proposed for air quality prediction, insufficient attention has been paid to investigating the performance of prediction models in different ranges of pollutant concentrations. Therefore, the purpose of the work was to study the concentration dependencies of the accuracy of forecasting the atmospheric pollution level with PM2.5 particles using the Long Short-Term Memory (LSTM), Extreme Gradient Boosting (XGBoost) and Random Forest methods based on the data obtained by the developed air quality monitoring system.

Methods and means of implementation

A dataset containing over 250 thousand records obtained during the year by the air pollution research system proposed in [18] was used for the analysis. This monitoring system was implemented according to the Internet of Things (IoT) paradigm, which involves the exchange of data between various sensors and computer systems in automatic mode using wireless telecommunications and standard communication protocols [19].

Air pollution monitoring was carried out using the HM-3301 sensor. The sensor is characterized by high accuracy and the ability to continuously determine the concentration of solid particles larger than 0.3 μm in real time due to used technology of laser beam scattering. Additionally, current values of the air temperature and relative humidity were measured using a DHT22 sensor. The Arduino UNO R4 WiFi microcontroller was used as a platform for collecting, pre-processing and transmitting sensor data to a remote server. Besides, information from the web resource https://www.weatherapi.com about the direction and speed of the wind for a given area was used for the analysis.

Data pre-processing consisted of removing records with incorrect data and preparing the input dataset for training machine learning models. The resulting dataset was characterized by an uneven distribution of records on the particulate matter concentration, as shown in Fig. 1.

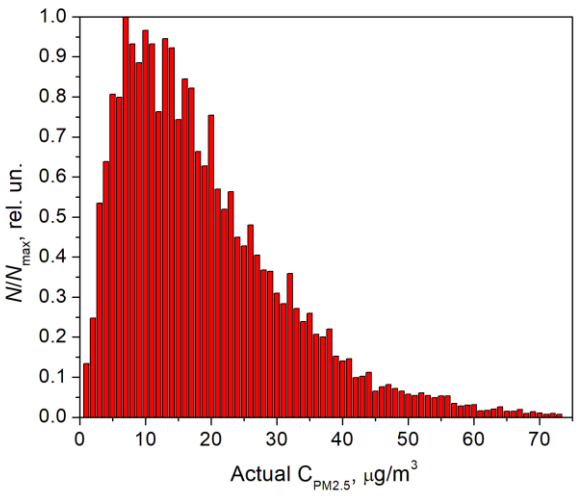


Fig. 1. Histogram of the distribution of records in the dataset on the concentration of the PM2.5 particulate matter in the air

In particular, the most frequently recorded values of the PM2.5 particle concentrations in the atmosphere were in the range of 7–14 μg/m^3. The number of records

with the concentration of PM2.5 more than 60 μg/m^3 was minimal. The average value of the level of air pollution with solid particles up to 2.5 μm in size was about 19 μg/m^3.

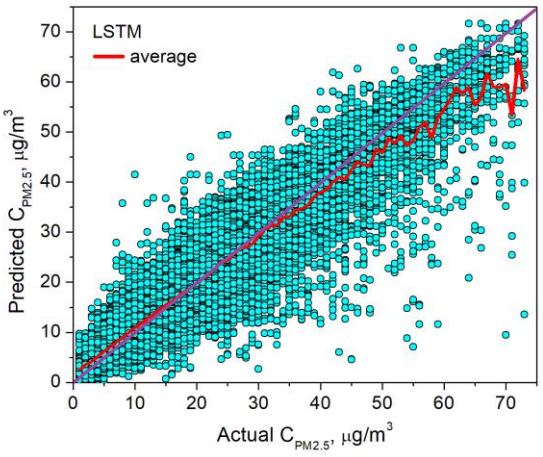
Temporal characteristics were represented using cyclic coding for months, days of the week and hours of the day. Seasonal patterns were taken into account in the additional attribute "season". The categorical feature characterizing wind direction was encoded using the one-hot encoding method. The prepared data was divided into training and test samples in a ratio of 80 and 20 %, respectively.

The data analysis was carried out using LSTM, XGBoost and Random Forest algorithms that demonstrate high accuracy and efficiency in predicting complex dependencies. In particular, due to the architectural features, the LSTM neural network can effectively detect and take into account patterns in time series and use them for forecasting. The strategies of the XGBoost and Random Forest ensemble methods not only provide high prediction accuracy for various datasets with complex nonlinear relationships but also enable the evaluation of the significance of individual features in the model. This assessment can potentially be used to identify the factors that most influence air pollution levels. The developed machine learning models were implemented in Python.

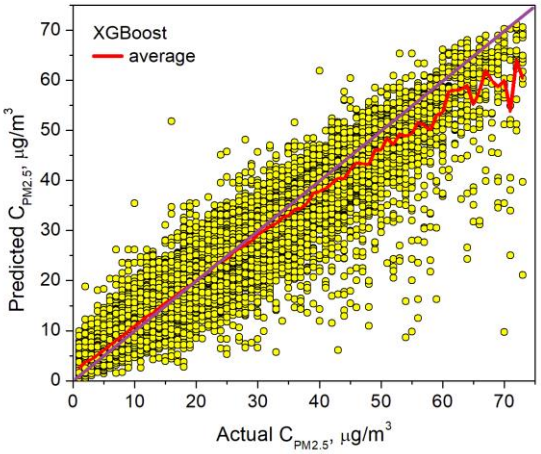
Results and discussion

The results of testing the proposed models for predicting the level of air pollution with PM2.5 particles demonstrate scatter plots that show the predicted values of the particulate matter concentration for each measured value of atmospheric pollution (Fig. 2).

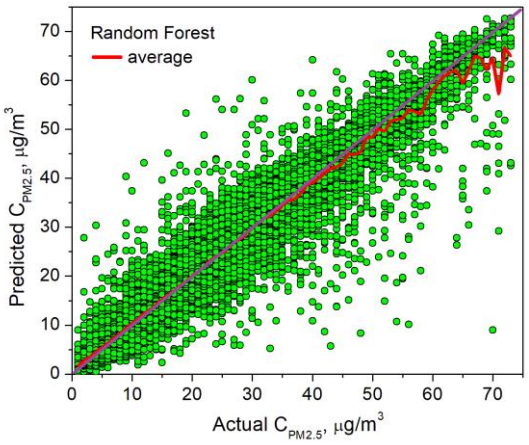
The obtained results are mostly grouped around the identity line for all three considered algorithms, which indicates a generally high prediction efficiency of the proposed models. The XGBoost and Random Forest methods demonstrate greater harmony of predicted values with the identity line but also allow significant deviations of some predictions for medium and high levels of air pollution. Such deviations occur mostly in the direction of underestimating the predicted values of the PM2.5 concentration in the atmosphere.



a)



b)



c)

Fig. 2. Scatter plots of the predicted values of the concentration of the PM2.5 particles in the air obtained by the LSTM (a), XGBoost (b) and Random Forest (c) methods

A similar trend is observed for the

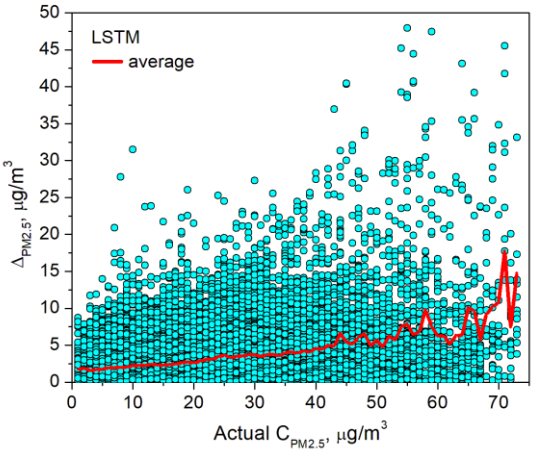
concentration dependencies of the averaged predicted values that show lower values from the identity line as the concentration of the solid microparticles increases (see Fig. 2). The observed pattern of discrepancies between predicted and measured values of the pollutant concentration is likely due to the random appearance of additional pollution sources unrelated to the meteorological conditions and spatio-temporal characteristics that are considered in the forecast.

The error in forecasting the level of air pollution by the proposed models also depends on the concentration of the PM2.5 particles. In particular, the Random Forest method is characterized by a lower absolute prediction error in most cases, as can be seen in Fig. 3. Despite this there are inaccuracies in some predictions in the entire concentration range. Forecasting by the LSTM and XGBoost methods almost does not allow significant deviations at low levels of atmospheric pollution, although generally larger values of the averaged absolute error are observed.

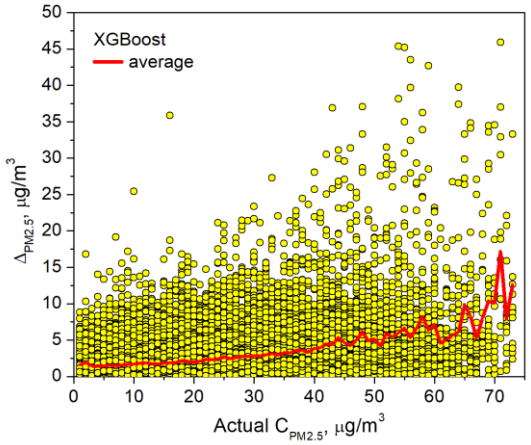
Analysis of the concentration dependencies of the relative error of air pollution forecasting makes it possible to claim that the model based on the Random Forest algorithm demonstrates the highest accuracy in the entire concentration range (Fig. 4). In general, the relative error of forecasting by this model is 6–9 % in the range of the average value of atmospheric pollution with PM2.5 particles (19 μg/m^3). The models based on the XGBoost and LSTM methods are characterized by errors of 9–11 and 11–14 %, respectively.

The proposed models demonstrate a quite high prediction accuracy for low levels of particulate matter pollution (up to 20 μg/m^3), as evidenced by the concentration dependencies of the dispersion of the predicted values shown in Fig. 5.

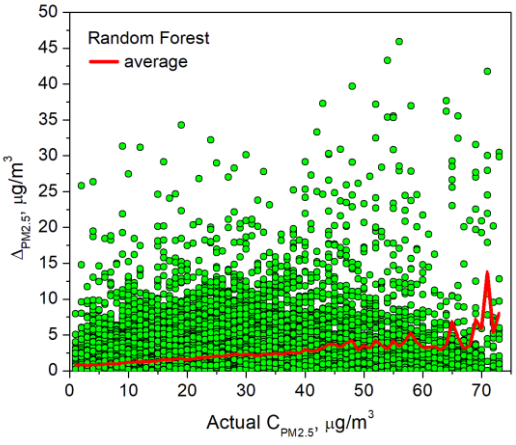
It is worth noting that a significant increase in the dispersion is observed with increasing the PM2.5 concentration for all implemented models. At the same time, the LSTM method demonstrates the least efficiency for high pollution levels. The increase in the variance of predicted values as well as absolute and relative errors with increasing air pollution may be due to the small number of records in the dataset with a high concentration of particulate matter (see Fig. 1).



a)



b)



c)

Fig. 3. Dependence of the absolute error of the air pollution prediction using the LSTM (a), XGBoost (b) and Random Forest (c) methods on the concentration of the PM2.5 particles in the air

The Mean Absolute Error (MAE), Mean Squared Error (MSE) and the coefficient of determination R^2 metrics, which provide an integral assessment of the predicted values in the entire range of the PM2.5 particle concentrations, were used to determine the effectiveness of the developed models for predicting the air pollution level based on meteorological and spatio-temporal data. The results of calculating the prediction accuracy using the LSTM, XGBoost and Random Forest methods according to the MAE, MSE and R^2 metrics are shown in Table 1.

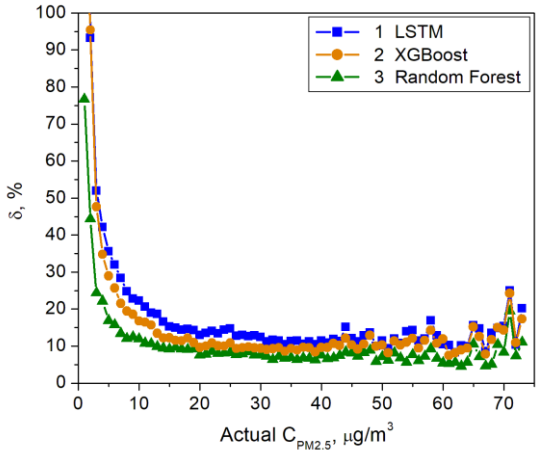


Fig. 4. Dependencies of the relative error of the air pollution prediction using the LSTM (1), XGBoost (2) and Random Forest (3) methods on the concentration of the PM2.5 particles

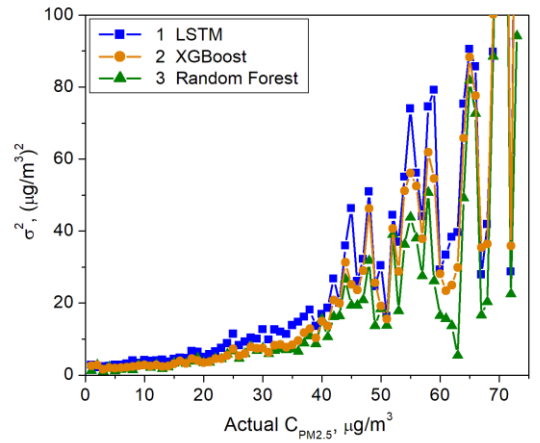


Fig. 5. Dependencies of the dispersion of the air pollution predicted values on the concentration of the PM2.5 particles using the LSTM (1), XGBoost (2) and Random Forest (3) methods

Analysis of the obtained results found that the model based on the Random Forest algorithm is characterized by the lowest MAE and MSE values as well as the highest coefficient of determination R^2. The XGBoost method demonstrates similar results. Slightly lower prediction accuracy is observed for the LSTM model.

Table 1. Air pollution level forecasting accuracy

Method	Metric		
	MAE	MSE	R^2
LSTM	2,86	18,13	0,89
XGBoost	2,28	13,02	0,92
Random Forest	1,65	9,09	0,94

Conclusions

The paper investigates the effectiveness of forecasting the level of air pollution using the LSTM, XGBoost and Random Forest methods based on historical values of the particulate matter concentration, temperature, air relative humidity, the direction and speed of the wind as well as temporal characteristics that take into account seasonality, days of the week and hours of the day. It was found that the proposed models successfully predict the PM2.5 content in the air at low and medium levels of pollution but underestimate the predicted values of particulate matter as their concentration increases.

Based on the analysis of the concentration dependencies of the absolute and relative errors and the dispersion of the predicted values, it was found that the Random Forest method outperforms the other two in terms of prediction accuracy although deviations are observed for some peak values. The LSTM and XGBoost methods almost do not allow significant errors and demonstrate good agreement between actual and predicted values of the PM2.5 concentration at low pollution levels. In general, the developed machine learning models effectively predict the concentration of the PM2.5 particle in the air as evidenced by high values of the coefficient of determination R^2.

References

1. Air quality in the world [Electronic resource]. - Mode of access: https://www.iqair.com/world-air-quality

2. Xing, Y. F., Xu, Y. H., Shi, M. H., & Lian, Y. X. (2016). The impact of PM2.5 on the human respiratory system. *Journal of Thoracic Disease*, 8(1), E69–74.

https://doi.org/10.3978/j.issn.2072-1439.2016.01.19

3. Vargas, J. E., Kubesch, N., Hernandéz-Ferrer, C., Carrasco-Turigas, G., Bustamante, M., Nieuwenhuijsen, M., & González, J. R. (2018). A systemic approach to identify signaling pathways activated during short-term exposure to traffic-related urban air pollution from human blood. *Environ. Sci. Pollut. Res.*, 25(29), 29572–29583. https://doi.org/10.1007/s11356-018-3009-8

4. Kan, H., Chen, R., & Tong, S. (2012). Ambient air pollution, climate change, and population health in China. *Environment International*, 42, 10–19. https://doi.org/10.1016/j.envint.2011.03.003

5. Kunzli, N., Jerrett, M., Mack, W. J., Beckerman, B., LaBree, L., Gilliland, F., Thomas, D., Peters, J., & Hodis, H. N. (2005). Ambient air pollution and atherosclerosis in Los Angeles. *Environ. Health Perspect.*, 113, 201–206. https://doi.org/10.1289/ehp.7523

6. Cieplak, T., Rymarczyk, T., & Tomaszewsk, R. (2019). A concept of the air quality monitoring system in the city of Lublin with machine learning methods to detect data outliers. *MATEC Web of Conferences*, 252, 03009. https://doi.org/10.1051/matecconf/201925203009

7. O'Leary, B., Reiners, J. J. Jr., Xu, X., & Lemke, L. D. (2016). Identification and influence of spatio-temporal outliers in urban air quality measurements. *Science of the Total Environment*, 573, 55–65. https://doi.org/10.1016/j.scitotenv.2016.08.031

8. Rukmani, P., Teja, G. K., & Vinay, M. S. (2018). Industrial monitoring using image processing, IoT and analyzing the sensor values using big data. *Procedia Computer Science*, 133, 991–997. https://doi.org/10.1016/j.procs.2018.07.077

9. Shankar, L., & Arasu, K. (2023). Deep learning techniques for air quality prediction: a focus on PM2.5 and periodicity. *Migration Letters*, 20(S13), 468–484. https://doi.org/10.59670/ml.v20iS13.6477

10. Kalajdjieski, J., Trivodaliev, K., Mirceva, G., Kalajdziski, S., & Gievska, S. (2023). A complete air pollution monitoring and prediction framework. *IEEE Access*, 11, 88730–88744. https://doi.org/10.1109/ACCESS.2023.3251346

11. Mokhtari, I., Bechkit, W., Rivano, H., & Yaici, M. R. (2021). Uncertainty-aware deep learning architectures for highly dynamic air quality prediction. *IEEE Access*, 9, 14765–14778. https://doi.org/10.1109/ACCESS.2021.3052429

12. Doreswamy, N., Harishkumar, K. S., Yogesh, K. M., & Gad, I. (2020). Forecasting air pollution particulate matter (PM2.5) using machine learning regression models. *Procedia Computer Science*, 171, 2057–2066. https://doi.org/10.1016/j.procs.2020.04.221

13. Qi, Y., Li, Q., Karimian, H., & Liu, D. (2019). A hybrid model for spatiotemporal forecasting of PM2.5 based on graph convolutional neural network and long short-term memory. *Science of the Total Environment*, 664, 1–10. https://doi.org/10.1016/j.scitotenv.2019.01.333

14. Mehmood, K., Bao, Y., Saifullah, Cheng, W., Khan, M. A., Siddique, N., Abrar, M. M., Soban, A., Fahad, S., & Naidu, R. (2022). Predicting the quality of air with machine learning approaches: Current research priorities and future perspectives. *Journal of Cleaner Production*, 379, 134656. https://doi.org/10.1016/j.jclepro.2022.134656

15. Liu, D., Lee, S., Huang, Y., & Chiu, C. (2020). Air pollution forecasting based on attention-based LSTM neural network and ensemble learning. *Expert Syst.*, 37(3), 1–12. https://doi.org/10.1111/exsy.12511

16. Xu, X., Tong, T., Zhang, W., & Meng, L. (2020). Fine-grained prediction of PM2.5 concentration based on multisource data and deep learning. *Atmospheric Pollution Research*, 11(10), 1728–1737. https://doi.org/10.1016/j.apr.2020.06.032

17. Yang, Y., Mei, G., & Izzo, S. (2022). Revealing influence of meteorological conditions on air quality prediction using explainable deep learning. *IEEE Access*, 10, 50755–50773. https://doi.org/10.1109/ACCESS.2022.3173734

18. Olenych, I., & Babiak, S. (2024). Automated air pollution research system. *Electronics and information technologies*, 26, 59–72, (in Ukrainian). https://doi.org/10.30970/eli.26.6

19. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of things: a survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys & Tutorials*, 17(4), 2347–2376. https://doi.org/10.1109/COMST.2015.2444095

The article has been sent to the editors 03.01.25.
After processing 07.02.25.
Submitted for printing 30.03.25.